Exploring Distinctive Features in Distant Supervision for Relation Extraction

Yang Liu, Shulin Liu, Kang Liu, Guangyou Zhou, and Jun Zhao

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{yang.liu,shulin.liu,kliu,gyzhou,jzhao}@nlpr.ia.ac.cn

Abstract. Distant supervision (DS) for relation extraction suffers from the noisy labeling problem. Most solutions try to model the noisy instances in the form of multi-instance learning. However, in the non-noisy instances, there may be noisy features which would harm the extraction model. In this paper, we employ a novel approach to address this problem by exploring distinctive features and assigning distinctive features more weight than the noisy ones. We make use of all the training data (both the labeled part that satisfies the DS assumption and the part that does not), and then employ an unsupervised method by topic model to discover the distribution of features to latent relations. At last, we compute the distinctiveness of features by using the obtained feature-relation distribution, and assign features weights based on their distinctiveness to train the extractor. Experiments show that the approach outperforms the baseline methods in both the held-out evaluation and the manual evaluation significantly.

Keywords: Relation Extraction, Distant Supervision, Distinctive Features.

1 Introduction

Relation Extraction is the task of extracting semantic relations between entity pairs given a set of sentences containing both entities. It gains much interest for its potential effects on constructing large scale knowledge bases and supporting many other applications like question answering [12], textual entailment [15] etc. Traditional supervised approaches for relation extraction [7][22] need to label training data, which is expensive and biased towards the domain of labeled data. Due to the problems of supervised approaches, an attractive paradigm called distant supervision (DS) [10] is employed. It automatically produces labeled training data by aligning entities in a knowledge base with relation facts (such as Freebase[1]) to sentences. However, it suffers from noisy labeled data which will bring poor extraction results. For example, in Figure 1, $r(e1, e2) =$ "$BornIn(Yao\ Ming, Shanghai)$" is a relation in the knowledge base. After automatic labeling, we get the sentences containing both $e1 =$ "$Yao\ Ming$" and $e2 =$ "$Shanghai$". The upper sentence truly express the relation $r =$ "$BornIn$" between two entities. However, the lower one does not. It is a noisy labeled sentence. In this paper, we focus to address the noisy labeling problem in DS for relation extraction.

[1] http://www.freebase.com/

R.E. Banchs et al. (Eds.): AIRS 2013, LNCS 8281, pp. 344–355, 2013.

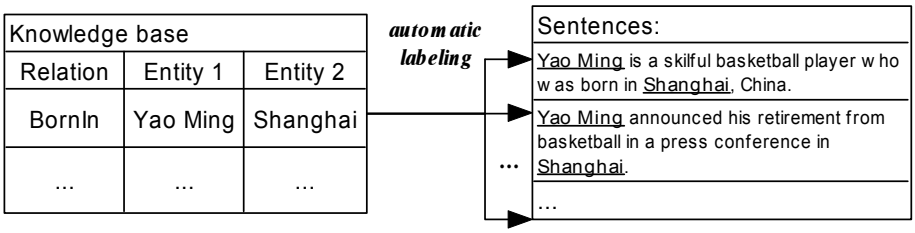

Fig. 1. Noise in training data by distant supervision. The first sentence is the correct labeling and the second one is incorrect.

To overcome the problem of noisy labeled data in DS, work in [13][9][14] attempted to model the noisy data with multi-instance learning methods. They assume that at least one of the sentences containing both $e1$ and $e2$ expresses $r(e1, e2)$. However, this at-least-one assumption can fail, for that Takamatsu et al. [16] showed 91.7% of entity pairs only have one labeled sentence in Wikipedia articles which do not fit for the multi-instance learning assumption. Moreover, they used binary features in their model. This setting will enforce some frequent indistinctive features. For example, labeled sentence "*Life of Pie, by Ang Lee, can be said to be the best...*" for *r(e1, e2) = "DirectorOf(Ang Lee, Life Of Pi)"* has one lexical feature "*e2 by e1*" where $e1$ and $e2$ are placeholders for two entities. However, it can also be found in the "*AuthorOf*" relation like the sentence "*One Hundred Years of Solitude, by Gabriel Garcia Marquez, is a novel that tells...*" for "*AuthorOf(Gabriel Garcia Marquez, One Hundred Years of Solitude)*". As a result, entity pairs of *AuthorOf* are probably mistaken for relation *DirectorOf*. Although the feature is from a positive sentence, it is still noisy. Binary features can not discriminate between the distinctive features and noisy ones.

In the paper, due to the deficiencies of the at-least-one assumption and the binary feature setting, we propose a novel approach to solve the noisy labeling problem. Instead of using binary features which take no difference to all features, we explore the distinctive features and assign distinctive features higher weight than the noisy ones. In this way, we do not use the at-least-one assumption as the multi-instance learning does and can solve the noisy feature problem in the non-noisy sentences that indeed express the target relations caused by the binary feature setting mentioned above.

Specifically, we employ a new method to calculate the distinctiveness of each feature. Our intuition is that the noisy features tend to appear in several different relations. It means that if a feature is used to indicate several different relations, it would be less distinctive. To obtain the feature-relation distribution, instead of only using the sentences labeled by a knowledge base with the DS assumption (**KB-matched instances**[2]) like previous work, we use the **united instances** that combine the KB-matched instances with the instances generated by entity pairs in training data but not in the knowledge base (**KB-not-matched instances**). For that, the KB-matched instances are only a small part of the training data, instances in which are biased to the relations used as labeling

[2] An instance consists of features of an entity pair extracted from all its labeled sentences.

sources and not sufficient to discover the distribution of features to latent relations. And then we employ a topic model to model the generating process of instances in united instances where the feature can be considered as "word", instances as "documents" and relations as "topics". After estimating the parameters, We can get the feature-relation distribution (or the "word-topic" distribution) via this model. After that, we compute the distinctiveness of each feature based on the feature-relation distribution, and assign features different weights according to their distinctiveness. Finally, we use these weighted features to train a classifier for discovering relations in new instances.

This paper mainly makes the following contributions:

- To solve the noisy training data problem, we propose a method to assign features different values according to features' distinctiveness to latent relations. We avoid the "at-least-one" assumption in multi-instance learning and can solve the noisy features extracted from positive labeled sentences such as "$e2$ *by* $e1$" in "*Life of Pie, by Ang Lee, can be said to be the best...*". To our best knowledge, little work has considered to weigh features for solving the noisy labeling problem in DS for relation extraction.
- To discover the probabilities of features belonging to latent relations, we model united instances combined KB-matched instances with KB-not-matched instances via a topic model and obtain the feature-relation distribution. Previous work mainly focused on the KB-matched instances, little work has tried to make use of unlabeled data (KB-not-matched instances), which do not satisfied the DS assumption.
- We conduct experiments to evaluate our method with Wikipedia articles and Freebase as the knowledge base. We compare our method with Mintz et al. [10] and the multi-instance learning approach of MULTIR [9]. The experimental results show that our method outperforms both methods.

The remainder of the paper is organized as follows. Section 2 introduces the related work. Section 3 describe the relation topic model. Section 4 introduces the method to weigh features. Section 5 illustrates our experiments and evaluation. Finally, we conclude our paper with the future work.

2 Related Work

Distant supervision (also known as weak supervision or self supervision) is used to a broad class of methods in information extraction which aims to automatically generate labeled data by aligning with data in knowledge bases. It is introduced by Craven and Kumlien [4] who used the Yeast Protein Database to generate labeled data and trained a naive-Bayes extractor. Bellare and McCallum [2] used BibTex records as the source of distant supervision. The KYLIN system in [18] used article titles and infoboxes of Wikipedia to label sentences and trained a CRF extractor aiming to generate infoboxes automatically. The Open IE systems TEXTRUNNER [21] and WOE [19] trained their extractors with the automatic labeled data from Penn Treebank and Wikipedia infoboxes respectively. Yao et al. [20] trained a CRF considering selectional preference constraints of entity types with weak supervision.

Our work was inspired by [10] which performed distant supervision for relation extraction. It used Freebase as the knowledge base to label sentences in Wikipedia as

training data and trained a logistic regression classifier to extract relations between entities. Distant supervision supplied a method to generate training data automatically, however it also bring the problem of noisy labeling. After their work, a variety of methods focused to solve this problem. Work in [16] predicted negative patterns using a generative model and remove labeled data containing negative patterns to reducing noise in labeled data. In [13][9][14], they proposed multi-instance learning methods with the assumption that at least one of the labeled sentences truly expressed their relation. However, this assumption does not fit for the entity pair with only one labeled sentence. We employ an alternative approach without the mentioned assumptions. Different from the previous work using binary features, we assign different weight to features according to their distinctiveness to target relations.

Algorithm 1. Unite KB-matched instances with KB-not-matched instances

Input:
The feature set of KB-mathced instances: $feat\_set(KB\_matched)$
KB-mathced instances: $Instances(KB\_matched)$
KB-not-mathced instances: $Instances(KB\_not\_matched)$
Output:
United instances: $Instances(U)$

```
tmpset=feat_set(KB_Matched)
pre_set_size = tmpset.size()
cur_set_size = pre_set_size
Instances(U) = Instances(KB_matched)
while pre_set_size != cur_set_size do
    for each instance in Instances(KB_not_matched) do
        for each feature in instance do
            if feature in tmpset then
                add instance to Instances(U)
                add features in instance to tmpset
                break
            end
        end
    end
    pre_set_size=cur_set_size
    cur_set_size = tmpset.size()
end
remove features with frequency below 5 from Instances(U)
return Instances(U);
```

3 Relation Topic Model

Aiming to discover the distribution of features to latent relations and due to , in this section, we first use features of KB-matched instances to combine with KB-not-matched instances, and then use a topic model for modeling features in united instances. At last we obtain the feature-relation distribution.

3.1 Generating United Instances

Given the training data set, previous work trained their models only using the sentences labeled by a knowledge base with the DS assumption (KB-matched instances2). However the KB-matched instances are only a small part of the training data, instances in them are biased to the relations used as labeling sources and not sufficient to discover the distribution of features to latent relations. As a result, we employ a method to unite KB-matched instances and KB-not-matched instances:

(a) First, after labeling with the DS assumption, we extract features for each entity pair from the matched sentences to obtain KB-matched instances[3]. The types of features are the same with the work [10].
(b) Second, we collect all entity pairs in training data except those generating KB-matched instances. The entity pairs are used as a source to match sentences from the training data with the DS assumption. And then we extract their features to obtain KB-not matched instances.
(c) Third, we united the two part of instances by features of KB-matched instances. We use the features of KB-matched instances to form a feature set, and then collect instances in KB-not-matched instances which contain at least one feature in the feature set. New features in the collected instances are added to the feature set. We iteratively do these steps until no new features can be found. The united instances consist of the KB-matched instances and the instances collected from KB-not matched instances(see Algorithm 1. for details).
(d) At last, we remove the features with frequencies below 5, for two reasons, one is that we consider features with low frequencies are distinctive so that we assign them high value directly, and the other is that their frequencies are too low to help discriminating the distinctiveness of frequent features by the topic model to be introduced next.

3.2 Modeling United Instances with Topic Model

Topic model (or LDA) [3] is a generative graphical model. It has achieved great success in finding the latent topic for documents. In this paper, we use it to model the generative process of each instance which has a set of features (Figure 2). We can consider a instance as an "document", features f in the "document" as "words" and latent relations r as latent "topics".

The generation process of topic model for each instance is as following:

1. Choose $N \sim$ Poisson(ξ).
2. Choose $\theta \sim$ Dirichlet(α).
3. Choose $\phi \sim$ Dirichlet(β).
4. For each of the N features f_n:
   (a) Choose a relation $r_n \sim$ Multinomial(θ)
   (b) Choose a feature f_n from $p(w_n|r_n, \phi_{r_n})$, a multinomial probability conditioned on the relation r_n.

[3] We mean the KB-matched instances as both the labeled positive instances and the negative instances (See Section 5.1).

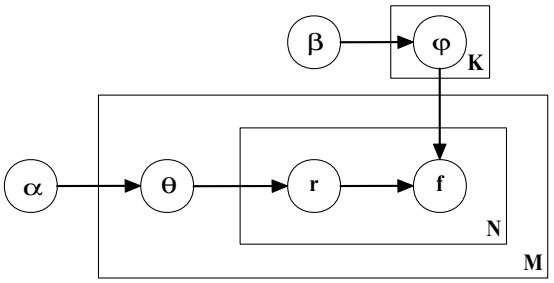

Fig. 2. Graphical model representation of topic model

Based on the generative graphical model depicted in Figure 2, the joint distribution of θ, $\mathbf{r}$ and $\mathbf{f}$ is given by:

$$p(\theta, \mathbf{r}, \mathbf{f}, \Phi|\alpha, \beta) = p(\Phi|\beta) \prod_{n=1}^{N} p(f_n|\phi_{r_n}) p(r_n|\theta) p(\theta|\alpha) \tag{1}$$

And the likelihood of an instance:

$$p(\mathbf{f}|\alpha, \beta) = \int \int p(\theta|\alpha) p(\Phi|\beta) \cdot \prod_{n=1}^{N} p(r_n|\theta, \Phi) d\Phi d\theta \tag{2}$$

Finally, taking the product of the likelihood of each instance, we get the probability of a relation corpus:

$$p(R|\alpha, \beta) = \sum_{m=1}^{M} p(\mathbf{f}|\alpha, \beta) \tag{3}$$

We estimate its parameters with Gibbs Sampling [8][11] and set number of relations as 50 and iteration times as 2000 in our experimetns. After estimation, we obtain a matrice $\Phi_{K \times N}$ representing the feature-relation distribution. The probability $p(f_i|r_k)$ of a feature f_i conditioned on a target relation r_k in $\Phi_{K \times N}$ is computed as follows:

$$p(f_i|r_k) = \phi_{k,i} = \frac{n_k^{(i)} + \beta_i}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v} \tag{4}$$

Where $n_k^{(i)}$ is the number of times that the ith feature is assigned to the kth relation. V is the size of features.

We will use the distribution to compute the distinctiveness of features in the next section.

4 Weighing Features by Their Distinctiveness

In this part, we use the obtained feature-relation distribution $\Phi_{K \times N}$ and the feature distribution in united instances to compute features' distinctiveness. Intuitively, if

features has equivalent probabilities among several latent relations, they are less distinctive than the ones which have significant probabilities in only one latent relation. We call it *clarity*. We measure the clarity for each feature by the following equation:

$$Clarity_{f_i} = \begin{cases} log_2 \frac{K \cdot max_{k \in \{1 \ldots K\}} p(f_i|r_k)}{\sum_{k=1}^{K} p(f_i|r_k)} \cdot \frac{1}{log_2 K}, \ K > 1 \\ \\ 1, \ K = 1, 0 \end{cases} \tag{5}$$

Where $p(f_i|r_k)$ is the probability of the ith feature f_i in the kth relation r_k from the relation-feature distribution $\Phi_{K \times N}$. If a feature is only observed once, its *clarity* is 1. If a feature can not be observed in features of the relation-feature distribution, its *clarity* is also 1. The reason is that the unobserved features are those with low frequencies, we consider they are less likely belonging to several relations.

Besides the clarity, intuitively, we think features with more information will tend to be less noisy. Our features are composed of lexical and syntactic pathes between two entities the same with [10]. More words in the pathes, more information the features will contain. For example, two feature "$e2$ *by* $e1$" and "$e2$ *directed by* $e1$" for the relation "$DirectorOf(Ang\ Lee,\ Life\ Of\ Pi)$", the latter one is more informative than the former one and it can better predict the target relation. And more, if a feature has a low frequency in united instances, it tends to be more specific to the relation containing this feature and be more predictable to this relation. As a result, less frequent features are more informative than more frequent ones. We measure features' *informativeness* with the following equation considering both the length and frequency mentioned above:

$$Informativeness_{f_i} = (\frac{len(f_i)}{max_{j \in \{1 \ldots n\}} len(f_j)})^{\alpha} \cdot (\frac{1}{freq(f_i)})^{\alpha} \tag{6}$$

In it, $len(f_i)$ denotes the number of words in feature f_i. $max_{j \in \{1 \ldots n\}} len(f_j)$ means the max number of words in features, $freq(f_i)$ is the frequency of the feature f_i in united instances. We use α ($0 < \alpha < 1$) to avoid values of features with high frequency or short length being too small. In the experiments, we set α as 0.25.

We compute the $distinctiveness$ of a feature by combining $clarity$ and $informativeness$. Based on the theory of Discriminative Category Mathcing (DCM) [6][1], we have the following equation, where $\sqrt{2}$ is a normalization factor:

$$Distinctiveness_{f_i} = \frac{Clarity_{f_i}^2 \cdot Informativeness_{f_i}^2}{\sqrt{Clarity_{f_i}^2 + Informativeness_{f_i}^2}} \cdot \sqrt{2} \tag{7}$$

We assign the $distinctiveness$ to each feature in KB-matched instances as its feature value, and then train a multi-class logistic classifier with Gaussian regularization as the extractor. Our extractor takes an entity pair and its feature vector as in put, and

Table 1. Nine relation types and their number of entity pairs in training and testing data labeled with the DS assumption

Relation Type	♯ in training data	♯ in testing data
location.country.administrative_divisions	1892	1441
location.location.contains	77120	50795
location.location.events	490	301
people.deceased_person.place_of_death	3591	1994
people.person.nationality	9717	5592
people.person.place_of_birth	7670	3785
film.film.directed_by	1501	856
film.film.written_by	1007	582
film.film.country	1404	873

return a relation name and its corresponding confidence score based on the probability it belongs to that relation. At last, we rank the extracting result based on their confidence to generate n most likely new relation instances and evaluate our method comparing to previous methods.

5 Experiments

5.1 Data

We conduct our experiments on articles of Wikepedia with Freebase as the knowledge base. We randomly sample 900,000 Wikipedia articles from Freebase Wikipedia Extraction (WEX)[4] data dump of 2012. In them, 600,000 articles are used as training data, and 300,000 are used as testing.

For preprocessing, we segment each article to sentences by XML tags in the WEX dump. To find entities in sentences, we first do NER tagging with Stanford NER [5]. We tag tokens into 5 categories: PERSON, ORGANIZATION, LOCATION, MISC and NONE where MISC means name entities not belonging to the first three categories. Adjacent name entities with the same NER tag are combined to one name entity. Then for entity pairs in sentences, we extract their features (see Section 3). The feature types are the same with [10] which mainly consist of lexical Part-Of-Speech (POS), name entity and syntactic features (paths between two entities in the dependency parsing tree). We use the Stanford POS tagger [17] to assign the Pos tags and Stanford parser[5] to parse the sentences. .

To distant supervision for relation extraction, we evaulate 9 of the most frequent relations in Freebase from three categories: people, location and film (see Table 1). To train our extractor, we need negative instances. As a result, we randomly sample 10% of the entity pairs that appear in the same sentence labeled by the DS assumption but are not contained in Freebase, and then use them to label negative instances.

[4] `http://wiki.freebase.com/wiki/WEX`

[5] `http://nlp.stanford.edu/software/stanford-dependencies.shtml`

5.2 Baselines

We compare our method ($PROP$) against two methods:

- $Mintz$: this method is implemented based on [10]. We use their aggregate feature setting to train a multi-class logistic regression classifier.
- $MULTIR$: this is the "at-least-one" model (a form of multi-instance learning) reported in [9]. It learns using a Perceptron algorithm. We use its released code[6] for our experiment.

5.3 Evaluation

Following the work in [10][9], we evaluate our method in two ways: the held-out evaluation and the manual evaluation. The held-out evaluation only compared the newly discovered relation instances against Freebase relation data, it would suffer from false negatives. Thus, besides the held-out evaluation, we further conduct the manual evaluation.

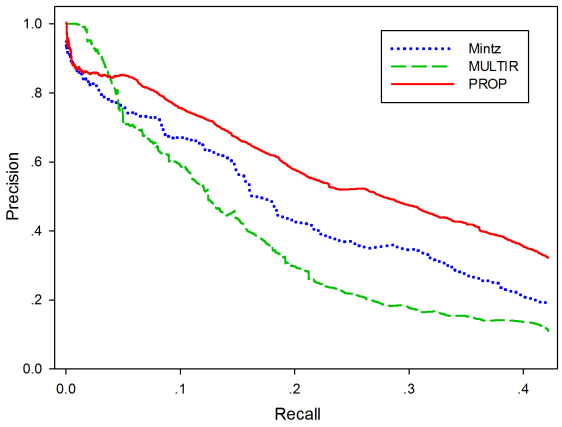

Fig. 3. Precision-recall curves in the held-out evaluation for three method: $Mintz$, $MULTIR$ and $PROP$

Held-Out Evaluation. In held-out evaluation, the extracted relation instances from testing data are automatically compared with those in Freebase. We rank the predicted relation instances by their confidences. Then we traverse this ranked list from high to low and measure precision and recall at each position.

Figure 2 shows the precision and recall curves for $Mintz$, $MULTIR$ and our proposed method $PROP$. At the head of the curves, $MULTIR$ outperforms the other two methods. However, it drops quickly below other two curves. $PROP$ is consistently outperforming $Mintz$ and it also achieve a better curve than $MULTIR$.

[6] `http://raphaelhoffmann.com/mr/`

Manual Evaluation. In manual evaluation, we remove the relation instances existing in Freebase and pick the top ranked 50 relation instances for each of the 9 relations. We manually label instances whether the relations indeed holds.

Table 2 shows the top 50 precisions of the 9 relations. Our approach $PROP$ outperforms $Mintz$ in 8 relations and outperforms $MULTIR$ in 4 relations. All the three methods fail in extracting the $film.film.country$ relation with no correct instance in its top 50 instances. Among the three methods, $PROP$ achieve the best average precision.

Table 2. Precision of manual evaluation of the top 50 ranked results for each relation

Relation Type	Precision		
	$Mintz$	$MULTIR$	$PROP$
location.country.administrative_divisions	0.82	0.64	**0.90**
location.location.contains	0.50	**0.98**	0.70
location.location.events	0.56	**0.64**	0.62
people.deceased_person.place_of_death	0.68	0.36	**0.72**
people.person.nationality	0.66	**0.92**	0.90
people.person.place_of_birth	0.68	0.90	**0.92**
film.film.directed_by	0.40	0.40	**0.58**
film.film.written_by	0.52	**0.74**	0.56
film.film.country	0	0	0
Average	0.54	0.62	**0.66**

Analysis. The experiment results show the advantage by exploring distinctive features and weighing features based on their distinctiveness. $Mintz$ used aggregate features which aggregates sentential binary faetures and $MULTIR$ used binary features. Their feature settings enforces some frequent noisy features in the labeled data generated with the distant supervision assumption like "$e2\ by\ e1$" for the $DirectorOf$ relation. Our method overcomes this problem.

$MULTIR$ learns a model driven by sentence-level features and aggregated sentence-level extracting results as a form of multi-instance learning. It alleviates the noisy labeling problem to some extent and achieves better results in some relations. However, because of the problem caused by the binary feature setting mentioned above, it performs quit bad in several relations. Taking the relation $people.deceased\_person.place\_of\_death$ as an example. We inspect its extracting result, it emphasizes the feature "$e1\ of\ e2$" like "$Barack\ Obama\ of\ Illinois$" which hurts its precision much.

The three methods failed in extracting the $film.film.country$ relation. The reason is that its automatically labeled data are in bad qualities. There are little specific information that can predicate this relation. The mistaken sentences are as follows: "$...to\ unite\ with [Czechoslovakia]_{e2}, [Harvard\ Ukrainian\ Research\ Institute]_{e1}$." and "$[Mohatta\ Palace]_{e1}\ -\ ([Karachi]_{e2})$." etc.

6 Conclusion

In this paper, we propose a new approach to address the noisy labeling problem in DS for relation extraction. Our method does not use the at-least-one assumption which can fail when there is only one labeled sentence, and it is able to handle the problem of noisy features in non-noisy instances. We explore distinctive features and assign distinctive features more weight than the noisy ones. We employ unsupervised topic model to discover feature-relation distribution in both KB-matched instances and KB-not-matched instances (united instances). And the feature-relation distribution are used to compute features' distinctiveness for weighing features. At last, we use the weighed features to train a classifier to discover relations of new instances.

In the future work, we will try to explore the features in all the training data that related to the labeled part of training data but not appeared in them. We expect they can help to improve the extracting performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61070106, No. 61272332 and No. 61202329), the National High Technology Development 863 Program of China (No. 2012AA011102), the National Basic Research Program of China (No. 2012CB316300) and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDD2 01201).

References

1. Akbik, A., Visengeriyeva, L., Herger, P., Hemsen, H., Löser, A.: Unsupervised discovery of relations and discriminative extraction patterns. In: Proceedings of the 24th International Conference on Computational Linguistics, pp. 17–32 (2012)
2. Bellare, K., McCallum, A.: Learning extractors from unlabeled text using relevant databases. In: Sixth International Workshop on Information Integration on the Web (2007)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
4. Craven, M., Kumlien, J., et al.: Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg, Germany, pp. 77–86 (1999)
5. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics (2005)
6. Fung, G.P.C., Yu, J.X., Lu, H.: Discriminative category matching: Efficient text classification for huge document collections. In: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2003, pp. 187–194. IEEE (2002)
7. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 427–434. Association for Computational Linguistics (2005)
8. Heinrich, G.: Parameter estimation for text analysis (2005), `http://www.arbylon.net/publications/text-est.pdf`

9. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 541–550 (2011)
10. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, pp. 1003–1011. Association for Computational Linguistics (2009)
11. Phan, X.-H., Nguyen, L.-M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100. ACM (2008)
12. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 41–47. Association for Computational Linguistics (2002)
13. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 148–163. Springer, Heidelberg (2010)
14. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 455–465. Association for Computational Linguistics (2012)
15. Szpektor, I., Tanev, H., Dagan, I., Coppola, B., et al.: Scaling Web-based aquisition of entailment relations. PhD thesis, Tel Aviv University (2005)
16. Takamatsu, S., Sato, I., Nakagawa, H.: Reducing wrong labels in distant supervision for relation extraction. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, pp. 721–729. Association for Computational Linguistics (2012)
17. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 173–180. Association for Computational Linguistics (2003)
18. Wu, F., Weld, D.S.: Autonomously semantifying wikipedia. In: Proceedings of the sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 41–50. ACM (2007)
19. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118–127. Association for Computational Linguistics (2010)
20. Yao, L., Riedel, S., McCallum, A.: Collective cross-document relation extraction without labelled data. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1013–1023. Association for Computational Linguistics (2010)
21. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: open information extraction on the web. In: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 25–26. Association for Computational Linguistics (2007)
22. Zhou, G., Zhang, M., Ji, D.H., Zhu, Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information (2007)